The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation

Ahmed Hamdi, Rahma Boujelbane, Nizar Habash, Alexis Nasr

The Effects of Factorizing Root and Pattern Mapping in Translating between Tunisian Arabic and Standard Arabic

Ahmed Hamdi[1] **Rahma Boujelbane**[1,2] **Nizar Habash**[3] **Alexis Nasr**[1]
(1) Laboratoire d'Informatique Fondamentale de Marseille, Aix-Marseille University
(2) Multimedia, InfoRmation Systems and Advanced Computing Laboratory
(3) Center for Computational Learning Systems Columbia University
{ahmed.hamdi,rahma.boujelbane,alexis.nasr}@lif.univmrs.fr
habash@ccls.columbia.edu

Abstract

The development of natural language processing tools for dialects faces the severe problem of lack of resources. In cases of diglossia, as in Arabic, one variant, Modern Standard Arabic (MSA), has many resources that can be used to build natural language processing tools. Whereas other variants, Arabic dialects, are resource poor. Taking advantage of the closeness of MSA and its dialects, one way to solve the problem of limited resources, consists in performing a translation of the dialect into MSA in order to use the tools developed for MSA. We describe in this paper an architecture for such a translation and we evaluate it on Tunisian Arabic verbs. Our approach relies on modeling the translation process over the deep morphological representations of roots and patterns, commonly used to model Semitic morphology. We compare different techniques for how to perform the cross-lingual mapping. Our evaluation demonstrates that the use of a decent coverage root+pattern lexicon of Tunisian and MSA with a backoff that assumes independence of mapping roots and patterns is optimal in reducing overall ambiguity and increasing recall.

1 Introduction

The Arabic language has many variants. Modern Standard Arabic (MSA) is one of them. It is the official language of all Arab countries. However, MSA is the native language of no Arabic speakers. It is used for education, printed and spoken media. There exists also a variety of Arabic dialects which are the native languages of Arabic speakers. Unlike MSA, Dialectal Arabic (DA) varieties are only spoken. Therefore, there is no standard orthographic conventions (Habash, 2010; Habash et al., 2012b).

Most of the Arabic natural language processing (NLP) resources are built in order to process MSA. Very few works on processing dialects have been established, and mainly for Egyptian, Iraqi and Levantine Arabic. In this work, we focus on the Tunisian Arabic dialect (TUN), an important yet less studied Arabic dialect. We propose to transform it into a form that is close to MSA by using morphological analysis and generation in order to take advantage of MSA NLP tools. Our approach relies on modeling the translation process over the deep morphological representations of roots and patterns, commonly used to model Semitic morphology. We compare different techniques for how to perform the cross-lingual mapping. Our evaluation demonstrates that the use of a decent coverage root+pattern lexicon of Tunisian and MSA with a backoff that assumes independence of mapping roots and patterns is optimal in reducing overall ambiguity and increasing recall.

The paper is organized as follows. We first present some related work in the next section. Section 3 discusses similarities and differences between MSA and TUN verbal morphology. In Section 4, we describe different tools that are used throughout this work. Section 5 evaluates our system.

2 Related Work

A limited amount of work has been done on building DA resources and tools, and mainly for Egyptian, Iraqi and Levantine Arabic. Maamouri et al. (2004b) presented a transcription corpus with its design principles, development tools and guidelines for speech recognition research. Habash et al. (2012b) developed a conventional orthography for dialectal Arabic (CODA) designed for developing computational models of Arabic dialects. CODA was used in the design of a morphological analyzer for Egyptian Arabic (Habash et al., 2012a), as well as a morphological disambiguation system for Egyptian Arabic (Habash et al., 2013) and a system for normalizing spontaneous orthography (Eskander et al., 2013). A morphological analyzer and generator for Arabic dialects (MAGEAD) was also developed for MSA and Levantine Arabic (Habash et al., 2005; Habash and Rambow, 2006; Altantawy et al., 2010; Altantawy et al., 2011). Al-Sabbagh and Girju (2010) described an approach of mining the web to build a DA-to-MSA lexicon. Riesa and Yarowsky (2006) presented a supervised algorithm for online morpheme segmentation on DA that cut machine translation out-of-vocabulary (OOV) words by half. Zbib et al. (2012) demonstrated an approach to cheaply obtaining DA-English data using crowd-sourcing.

Several researchers have considered the idea of exploiting existing MSA rich resources to build tools for DA NLP. For example, in order to use MSA treebanks to parse Levantine Arabic, Chiang et al. (2006) compared three methods that rely on translating between MSA and Levantine. Abo Bakr et al. (2008) introduced a hybrid approach to transfer a sentence from Egyptian Arabic into MSA. Sawaf (2010), Salloum and Habash (2011) and Salloum and Habash (2013) converted DA into MSA using a dialectal morphological analyzer and various mapping rules. Salloum and Habash (2011)'s DA morphological analyzer (ADAM), was built by extending a MSA analyzer in a noisy fashion. Their goal was to maximize analyzability not correctness. Mohamed et al. (2012) described a method for translating disambiguated MSA to Egyptian Arabic using a rule-based system. Their system reduced OOVs and improved POS tagging accuracy.

In this paper, we explore a similar approach to previous efforts (Sawaf, 2010; Mohamed et al., 2012; Salloum and Habash, 2013) but using a well-motivated deep morphological representation based on the MAGEAD approach (Habash and Rambow, 2006). Our solution is bi-directional unlike previous efforts and we demonstrate our approach on Tunisian Arabic.

3 Morphology: MSA vs Tunisian Arabic

Many similarities and differences exist between MSA and TUN in every aspect of verbal morphology: cliticization, inflection and derivation.

3.1 Cliticization Morphology

Various particles, called clitics, attach to inflected words. Clitics are optional and do not change the core meaning of the verbs they attach to. There are two main differences in cliticization morphology between MSA and Tunisian. First, several MSA clitics change their form in Tunisian. For example, the MSA interrogative particle proclitic (prefixing clitic) +أ *Âa+*[1] becomes the enclitic (suffixing clitic) ش+ *+š*. Second, some MSA clitics become detached in TUN and vice versa. The MSA future particle proclitic +س *sa+* is realized as the autonomous particle باش *bAš* with TUN verbs. Inversely, indirect object pronouns are realized as enclitics in TUN verbs and not in MSA. The general structure of MSA and TUN verbs is represented in the following two regular expressions:

QST? CNJ? PRT? MSA_VERB PRN_D?

CNJ? PRT? TUN_VERB PRN_D? PRN_I? (NEG|QST)?

QST (question) is the interrogative particle, CNJ is either the conjunctions +و *w+* 'and' or +ف *f+* 'so'. PRT is the class of particle proclitics such as future, prepositional and negation particle. NEG is a negation enclitic specified for TUN used with a negation proclitic. PRN_D and PRN_I are the direct and indirect object pronouns, respectively.

3.2 Inflectional and Derivational Morphology

Arabic words are constructed using two kinds of morphological operations: templatic and affixational. Functionally, both operations are used inflectionally or derivationally (Habash, 2007). In templatic morphology, a typically triliteral root and a pattern combine to form a word's stem, which is then extended with prefixes and suffixes, e.g., the TUN verb ومانقارنوهاش *wmAnqArnuwhAš* 'and we do not compare her/it' can be analyzed as *w+mA+n-*$\{\frac{1A23}{\sqrt{qrn}}\}$*-uw+hA+š*, where 1A23 is the

[1]Arabic orthographic transliteration is presented in the HSB scheme (Habash et al., 2007): (in alphabetical order)
ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
A b t θ j H x d ð r z s š S D T Ď ς γ f q k l m n h w y
and the additional letters: ' ء, Â أ, Ǎ إ, Ā آ, ŵ ؤ, ŷ ئ, ħ ة, ý ى.

pattern, $\sqrt{qrn}$ the root, clitics are marked with '+' delimiter and affixes with '-' delimiter. MSA has a richer inflectional morphology than TUN. In fact, some MSA features such as nominal case and verbal mood do not exist in TUN. Furthermore, the MSA number values of singular, dual and plural are reduced to singular and plural. Masculine and feminine values of gender feature are not distinguished in TUN except for the third person singular. Patterns carry a general meaning, the MSA pattern Ai12a33, for example, denotes the change of state. This pattern is not used in TUN and Tunisians express the state change by using the pattern 12A3 which not exists in MSA. Furthermore, some MSA patterns are not defined in TUN and vice versa.

4 Tools and Resources

Our architecture relies on the morphological processing tool MAGEAD and on a transfer lexicon.

4.1 MAGEAD

MAGEAD (Habash and Rambow, 2005) is a morphological analyzer and generator for the Arabic language family (MSA and Arabic dialects). MAGEAD relates (bidirectionally) a lexeme and a set of linguistic features to a surface word form through a sequence of transformations. In a generation perspective, the features are translated to abstract morphemes which are then ordered, and expressed as concrete morphemes. The concrete templatic morphemes are interdigitated and affixes added, finally morphological and phonological rewrite rules are applied.

4.1.1 Lexeme and Features

Morphological analyses are represented in terms of a lexeme and features. The lexeme is defined as a root, a morphological behavior class (MBC). We use as our example the surface form ازدهرت *Aizdaharat* 'she flourished'. The MAGEAD lexeme-and-features representation of this word form is as follows:

(1) Root:zhr MBC:verb-VIII POS:V PER:3 GEN:F NUM:SG ASPECT:PERF

4.1.2 Morphological Behavior Class

An MBC maps sets of linguistic feature-value pairs to sets of abstract morphemes. For example, MBC verb-VIII maps the feature-value pair ASPECT:PERF to the abstract root morpheme [PAT_PV:VIII], which in MSA corresponds to the concrete root morpheme V1tV2V3, while the MBC verb-II maps ASPECT:PERF to the abstract root morpheme [PAT_PV:II], which in MSA corresponds to the concrete root morpheme 1V22V3. MBCs are defined using a hierarchical representation with non-monotonic inheritance. The hierarchy allows to specify only once those feature-to-morpheme mappings for all MBCs which share them. For example, the root node of MSA MBC hierarchy is a word, and all Arabic words share certain mappings, such as that from the linguistic feature conj:w to the clitic w+. This means that all Arabic words can take a cliticized conjunction. Similarly, the object pronominal clitics are the same for all transitive verbs, no matter what their templatic pattern is.

4.1.3 MAGEAD Morphemes

To keep the MBC hierarchy variant-independent, a variant-independent representation of the abstract morphemes (AMs) that the MBC hierarchy maps to have been chosen. The AMs are then ordered into the surface order of the corresponding concrete morphemes. The ordering of AMs is specified in a variant-independent context-free grammar. At this point, our example (1) looks like this:

(2) [Root:zhr][PAT_PV:VIII][VOC_PV:VIII-act] + [SUBJSUF_PV:3FS]

Note that the root, pattern, and vocalism are not ordered with respect to each other, they are simply juxtaposed. The '+' sign indicates the ordering of affixational morphemes. Only now are the AMs translated to concrete morphemes (CMs), which are concatenated in the specified order. Our example becomes:

(3) <zhr,V1tV2V3,iaa> + at

Simple interdigitation of root, pattern and vocalism then yields the form iztahar+at.

4.1.4 MAGEAD Rules

MAGEAD uses two types of rules. Morphophonemic/phonological rules map from the morphemic representation to the phonological and orthographic representations. Orthographic rules rewrite only the orthographic representation. For our example, we get /izdaharat/ at the phonological level (as opposed to /iztaharat/). Using standard MSA diacritized orthography, our example becomes *Aizdaharat*. Removing the diacritics turns this into the more familiar *Azdhrt*. We follow

(Kiraz, 2000) in using a multi-tape representation. MAGEAD extend the analysis of Kiraz by introducing a fifth tier. The five tiers are used as follows: Tier 1: pattern and affixational morphemes; Tier 2: root; Tier 3: vocalism; Tier 4: phonological representation; Tier 5: orthographic representation. In the generation direction, tiers 1 through 3 are always input tiers. Tier 4 is first an output tier, and subsequently an input tier. Tier 5 is always an output tier.

4.1.5 From MSA to Tunisian

We adapted MAGEAD to process TUN verbs. Our effort concentrated on the orthographic representation. Changes concerned only the representation of linguistic knowledge, leaving the processing engine unchanged. We modified the MBC hierarchy, adding one MBC, removing three and editing five. The AM ordering has been modified and a new AM has been added for indirect object. The mapping from AMs to CMs and the definition of rules, which are variant-specific, are obtained from a linguistically trained native speaker. Furthermore, we needed to change some morphophonemic rules. In MSA, for example, the gemination[2] rule, allows deleting the vowel between the second and the third radical if it is followed by a suffix starting with a vowel: compare مددت *madad+tu* 'I extended' with مدّت *mad∼+at* 'she extended' (NOT *madad+at*). In Tunisian, in contrast, gemination always happens, independently of the suffix: مدّيت *mad∼+iyt* 'I extended' and مدّت *mad∼+it* 'she extended'. Many other rule changes were needed for TUN. For example, the first root radical becomes a long vowel in the imperfective aspect when it corresponds to ء ' (*hamza/glottal stop*) (يأكل *yÂkl* becomes ياكل *yAkl* 'he/it eats'). On the other hand, verbs whose root ends with ء ', behave the same way as verbs whose final root radical ي *y* in the perfective aspect. For example, roots of TUN verbs بدينا *bdynA* 'we started' and رمينا *rmynA* 'we threw' are respectively ب د ء *b d '* and ر م ي *r m y*. More details are discussed in Hamdi et al. (2013).

4.2 Root and Pattern Lexicon

Our lexicon is made of pairs of the form (P_{MSA}, P_{TUN}) where P_{MSA} and P_{TUN} are themselves pairs made of a root and an MBC. Its development was based on the Arabic Tree Bank (ATB) (Maamouri et al., 2004a) which contains 29,911 verb tokens. In order to extract the lemmas and the roots of these verbs, we used the morphological analyzer ElixirFM (Smrž, 2007) which extracts the lemma and the root of MSA inflected forms.[3] Then, each token of MSA lemma was translated by a Tunisian native speaker. At this point, lexicon entries are composed of a lemma and a root on the MSA side but only a lemma on the TUN side. We then associated to every entry an MBC (on the MSA side) and an MBC and a root (on the TUN side). In 81.49% of cases, we identified an Arabic existing root for TUN verbs. When there was no root for a given lemma, we used a deductive method to create a new one. Indeed, given the equation root + pattern = lemma, when we have a lemma and a pattern, it is possible to deduce a root. Using this process, we defined 100 new specific Tunisian roots.

In its current state, the lexicon contains 1,638 entries. The TUN side contains 920 distinct pairs and the MSA side 1,478 distinct pairs. As expected, the ambiguity is more important in the TUN $\rightarrow$ MSA sense. On average, a TUN pair corresponds to 1.78 MSA pairs, 1.11 in the opposite direction. The maximum ambiguity is equal to four in the MSA $\rightarrow$ TUN direction and sixteen in the opposite direction. More will be said about ambiguity in Section 5.

A sample of the lexicon appears in Table 1. The MBC indicates the pattern and in some cases the short vowels of the second root radical in the perfective and the imperfective aspects since they could change from verb to other. As shown in the table, a MSA MBC could be mapped to many TUN MBCs and vice versa.

Two by-products can be built form the lexicon, a root lexicon and a pattern correspondence table, both described below.

4.2.1 Root Lexicon

The root lexicon is made of pairs of the form (r_{MSA}, r_{TUN}), where r_{MSA} is an MSA root and r_{TUN} is a TUN root. The root lexicon contains 1,329 entries. The MSA side contains 1,050 dis-

[2]The second and the third root radical are identical.

[3]We did not use MAGEAD to perform the root extraction because the work on the lexicon had already started independently. MAGEAD for MSA, whose lexicon is based on the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) – just like ElixirFM, could have been used in principle.

MSA		TUN		English
Root	**MBC / Pattern**	**Root**	**MBC / Pattern**	**Gloss**
Smt	1-aa / 1a2a3	skt	1-ii / 12i3	'to be silent'
Hlq	1-aa / 1a2a3	Hjm	2-ii / 1a22i3	'to cut hair'
rtb	2 / 1a22a3	nZm	2-ii / 1a22i3	'to rank'
Hlq	2 / 1a22a3	Tyr	1-a / 12a3	'to fly'
xSm	3 / 1A2a3	ςrk	3-ii / 1A2i3	'to dispute'
dhm	3 / 1A2a3	hjm	1-ii / 12i3	'to attack'
bhr	4 / Aa12a3	ςjb	1-ii / 12i3	'to amaze'
xfy	4 / Aa12a3	xby	2-ai / 1a22a3	'to hide'
ršf	5 / ta1a22a3	ršf	5-ii / t1a22i3	'to savor'
ςjb	5 / ta1a22a3	bht	1-ii / 12i3	'to be surprised'
šjr	6 / ta1A2a3	ςrk	6 / t1A2i3	'to fight'
ςfy	6 / ta1A2a3	bry	1-aa / 12a3	'to be cured'
xfD	7 / Ain1a2a3	nqS	1-uu / 12u3	'to decrease'
sHb	7 / Ain1a2a3	bTl	2-ii / 1a22i3	'to step down'
nhy	8 / Ai1ta2a3	kml	1-ii / 12i3	'to be end'
Hdn	8 / Ai1ta2a3	Hml	2-ii / 1a22i3	'to hold'
dςy	10 / Aista12a3	ςdy	10 / Aista12a3	'to invite'
wfy	10 / Aista12a3	kml	2-ii / 1a22i3	'to complete'

Table 1: A sample TUN-MSA lexicon. The pattern provided is the form used with 3rd masculine singular perfective inflection. It is only presented for illustrative reasons to exemplify and highlight differences between TUN and MSA MBCs.

tinct roots and the TUN side 646 ones. 519 entries are composed of the same root on both sides. As in the root and pattern lexicon, the ambiguity is higher in the TUN → MSA direction. On average, a TUN root is paired with 2.06 MSA roots. In the opposite direction, this figure is equal to 1.26.

4.2.2 Pattern Correspondence Table

The pattern correspondence table indicates, for a pattern in MSA or TUN, the most frequent corresponding pattern in the other side. The pattern correspondence table is itself built on a pattern correspondence matrix, which is represented in Table 2. Each line of the matrix corresponds to a MSA pattern and each column to a TUN pattern. The matrix reads as follow, MSA pattern 1, for example, corresponds in 434 times to TUN pattern 1, 98 times to TUN pattern 2, and so on.

This matrix reveals several interesting facts. First, all patterns are not present in MSA or TUN in our lexicon. Pattern 9, for example is absent both in MSA and TUN and patterns 4 and 7 are absent on the TUN side. Second, there is a general tendency to keep the same pattern on the source and target sides of a lexicon entry. This is represented in the matrix by the fact that figures on the diagonal (in bold face) usually are the highest figure of both their line and column (the only exception is pattern 8). When a pattern does not exist in TUN, it is usually mapped to pattern 1.

	TUN							
		1	2	3	5	6	8	10
MSA	1	**434**$^*_*$	98	10	15		2	
	2	39	**298**$^*_*$	2	2	2		2
	3	24	19	**56**$^*_*$		2		
	4	69	**118***	4	6			
	5	26	16	2	**88**$^*_*$			3
	6	18	14	2	7	**26**$^*_*$		
	7	**13***	7	2				
	8	**41***	24	5	16	4	**18**$_*$	
	10	17	24	2	3			**31**$^*_*$

Table 2: Pattern correspondence matrix. Bolded cells are either the highest counts when translating from TUN to MSA or from MSA to TUN. X^* indicates highest count from MSA to TUN; and X_* indicates highest count from TUN to MSA.

The extraction of the pattern correspondence tables form the pattern correspondence matrix is straightforward: it consists in selecting for every pattern in the source side the most frequent pattern for the target side. It is interesting to note that is some cases, the most frequent pattern clearly dominates the other patterns, as it is the case for pattern 2 in MSA. In other cases, the tendency is not clear, as in pattern 4 in MSA.

Overall, the matrix tells us that selecting a target root and a target pattern are not independent processes. In other words, the root and pattern lexicon contains more information than the root lexicon along with the pattern correspondence table. We will experimentally quantify, in Section 5, the influence of making such an independence hypothesis.

5 Evaluation

The process of translating a source verbal form to a target verbal form proceeds in three main steps: morphological analysis using MAGEAD for the source language, followed by lexical transfer of roots and MBCs and finally, morphological generation of target verbal forms. All of these steps are reversible.

The whole process contains two sources of ambiguity: the analysis can create multiple (root, MBC) pairs and the lexicon may propose for an input pair many target pairs.

As we mentioned in the introduction, the goal of this work is not translation for TUN to MSA but generating from a TUN text an approximation of MSA, so that MSA NLP tools, such as morpho-syntactic taggers or parsers can be applied to this new form of text with acceptable results. The experiments described here provide only a partial evaluation, they allow to measure the proportion of cases in which the correct MSA form is generated given a TUN form.

The evaluation process is faced with the problem of lack of written resources for dialects. To overcome this problem, we used a book by Dhouib (2007) which is a Tunisian theater piece. 1500 tokens of TUN verbal forms were identified and translated in context to MSA by two Tunisian native speakers. At the end of this process, 1500 pairs were produced. This set was divided into two equal parts. The first was used as a development set and the second as a test set. Two standard metrics were used to evaluate the process: recall, which indicates the proportion of cases where the correct target form was produced; and ambiguity, which indicates the number of target forms produced on average for an input. The development set allowed us to fill some gaps in MAGEAD and enrich our lexicon.

We conducted the evaluation on undiacritized verbal forms since most of written Arabic is undiacritized. Without neither morphological nor lexical transfer, recall reaches 30.93% on tokens and 29.44% on types[4] but ambiguity is still at 1.0. This experiment gives the ratio of identical undiacritized TUN and MSA verbal forms in the test set.

In the following four sections, we present a series of experiments with different ways of realizing the transfer especially with respect to factorizing roots and patterns.

5.1 Pattern Correspondence Table

The most simple transfer process that we have experimented consists in leaving the source root unchanged and selecting the target pattern by a pattern correspondence table lookup. This experiment corresponds to the situation in which we do not have at our disposal a transfer lexicon. Since pattern is defined as a superset of MBCs, the target pattern maps to many target MBCs, each of them is associated to the target root and features to form the input of the morphological generator. We have chosen to build a correspondence pattern table instead of a correspondence MBC table for two main reasons : first, evaluations are made in an undiacritized set of verbs. Second, patterns carry a general meaning which can be a way to match MSA with TUN patterns. A block diagram of the process is presented in Figure 1 and the result of the experiment can be found in Table 3.

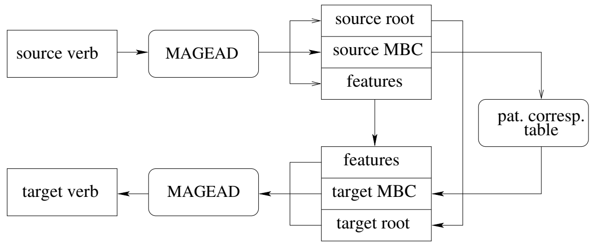

Figure 1: Translation process of source verbal form to target verbal form using a pattern correspondence table

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	47.74	43.40	39.41	37.61
MSA → TUN	52.55	48.05	5.89	7.12

Table 3: Recall and ambiguity on test set using pattern correspondence table

Table 3 shows two interesting features. First, the recall is quite low, around 50%. Keeping the source root is therefore a very rough approximation of the target variant. Second, the ambiguity is much higher in the TUN→MSA direction. This is due to the fact that TUN forms are morphologically more ambiguous than MSA forms. On average, a TUN form has 24.05 different analyses while MSA forms has on average 10.21 analyses. As mentioned in Section 3 MSA has a richer inflectional morphology than TUN, however our system used the same features for TUN and MSA analysis. Consequently, when a feature does not exist on TUN side, it produces many identical analysis with different values of this feature and generates subsequently many MSA verbal forms.

The same experiment was done using two target patterns instead of one (see Table 4). Table 4 shows a slight increase in recall. It rises on tokens to 51.65% in the TUN→MSA direction and 53.96% in the other direction. However, the ambiguity becomes higher, the process produces about

[4]Types are unique instances of tokens.

70 MSA verbs on average for a TUN token.

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	51.65	48.23	66.98	64.69
MSA → TUN	53.96	50.87	9.81	10.68

Table 4: Recall and ambiguity on test set using pattern correspondence table

5.2 Root Lexicon and Pattern Correspondence Table

In this experiment, the target pattern is selected as before by a lookup in the pattern correspondence table but the target roots are selected by a root lexicon lookup. This new setting was devised in order to increase the recall by better modeling root modification. The block diagram of the new setting appears in Figure 2 and the results on test set in Table 5 and 6.

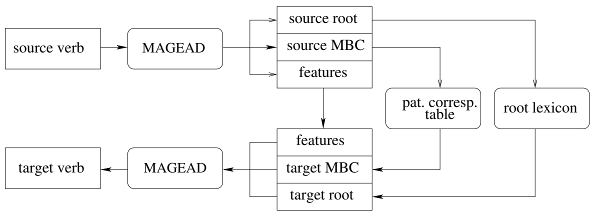

Figure 2: Translation process of source verbal form to target verbal form using a root lexicon and a pattern correspondence table

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	68.98	66.56	74.37	72.89
MSA → TUN	72.37	71.60	13.70	14.52

Table 5: Recall and ambiguity on test using a root lexicon and a pattern correspondence table

As expected, Table 5 shows a significant improvement of the recall. Ambiguity has also increased, this is due to the fact that a source root can map to several target roots: on average 2.06 in the TUN→MSA direction and 1.26 in the opposite direction.

Using the two most frequent target patterns from the pattern correspondence table, the translation process gives the highest recall and ambiguity, as shown in Table 6. In the MSA→TUN direction, recall rises to 86.12% on tokens and 81.77% in the inverse direction. The downside of this process is the ambiguity which becomes more then 100 in the TUN→MSA direction.

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	81.77	80.66	126.44	122.45
MSA → TUN	86.12	84.97	21.92	22.56

Table 6: Recall and ambiguity on test using a root lexicon and a pattern correspondence table

5.3 Root and Pattern Lexicon

In the preceding experiment, target roots and target patterns are translated independently and paired to compose the input of the morphological generator. But, as mentioned in Section 1, target root selection and target pattern selection are not independent processes: two source (root, pattern) pairs, sharing a common pattern can select different target patterns. In such cases the preceding method will give birth to incorrect (root, pattern) pairs and, eventually, incorrect verbal forms. In this experiment, target roots and patterns are selected together by a root and pattern lexicon access. The new process is represented in Figure 3 and results appear in Table 7.

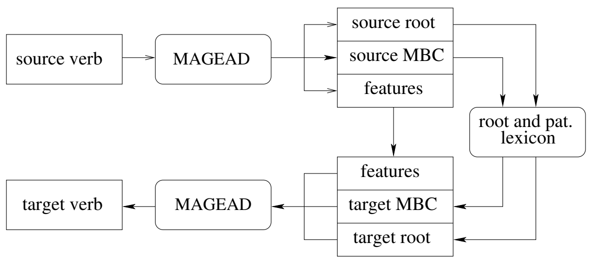

Figure 3: Translation process of source verbal form to target verbal form using a root and pattern lexicon

	recall		ambiguity	
	tokens	types	tokens	types
TUN → MSA	76.43	74.52	26.82	25.57
MSA → TUN	79.24	75.10	1.47	3.10

Table 7: Recall and ambiguity on test using a root and pattern lexicon

Replacing the root lexicon and the pattern correspondence table by a root and pattern lexicon has

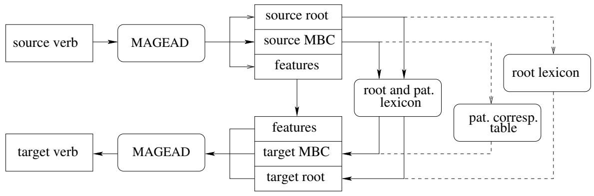

Figure 4: Translation process of source verbal form to target verbal form using a root and pattern lexicon with backoff on a root lexicon and a pattern correspondence table

a positive effect both on recall and ambiguity. The difference between the results of this experiment and the preceding one allows us to quantify the independence hypothesis of the root selection and the pattern selection we made in the preceding experiment.

The main weakness of this method is lexical coverage. We cannot expect to have a complete root and pattern lexicon and, sometimes, lexicon access fails. It is interesting at this point to mention the results of the same experiment on the development set. Recall that the verbal forms included in the development set have been used to populate the lexicon. As a consequence, a lexicon access never fails, and always produces the correct target (root, pattern) pair. The results of such an experiment, although artificial, allow to estimate an upper bound of such a method. In TUN $\rightarrow$ MSA direction, recall on tokens reaches 87.65% and in the inverse direction, it reaches 89.56%.

The reason why we did not reach 100% recall in this experiment is due to the fact that both MSA and TUN MAGEAD do not always produce the correct analysis, when used as an analyzer, or the correct form when used as a generator. An error analysis in the TUN $\rightarrow$ MSA direction showed that 21.8% of errors come from MSA MAGEAD and 78.2% from TUN MAGEAD. Most MAGEAD mistakes are due to morphological phenomena which have not been implemented yet, as quadriliteral verbs and the imperative form of defective verbs.[5]

5.4 Root and Pattern Lexicon with Backoff

In order to deal with low lexical coverage, we devised a variant of the preceding method which backs off, in cases of lexicon lookup failure, to the root lexicon and a the pattern correspondence table. The architecture of the system is shown in Figure 4, where the dotted lines represent the backoff path.

As Table 8 shows, this method increases recall significantly. This increase is itself the result of a better coverage. Ambiguity has also increased, this is due to the fact that when backing off, the transfer tends to be more ambiguous.

	recall		ambiguity	
	tokens	types	tokens	types
TUN $\rightarrow$ MSA	79.71	78.94	29.16	28.44
MSA $\rightarrow$ TUN	84.83	84.03	3.47	4.95

Table 8: Recall and ambiguity on test using a root and pattern lexicon with backoff on a root lexicon and a pattern correspondence table

6 Conclusion and Future Work

We presented a translation system between MSA and TUN verbal forms. This work is part of a wider project of translating Arabic dialects to an approximation of MSA. The results given by our system are about 80% recall in the TUN $\rightarrow$ MSA direction and 84% recall in the opposite direction. The translation process is highly ambiguous, in the MSA $\rightarrow$ TUN direction, the mean ambiguity is equal to 3.47 and reaches 29.16 in the opposite direction. A contextual disambiguation process is therefore necessary for such a process to be of practical use.

Future work will involve the development of a morphological model for nouns for TUN following the work of Altantawy et al. (2010), as well as a lexicon. In parallel we will work on the disambiguation of the TUN $\rightarrow$ MSA translations using a language model trained on a MSA corpus.

Acknowledgments

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Nos. HR0011-12-C-0014 and HR0011-12-C-0016. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA.

[5] Arabic defective verbs contain /w/ or /y/ in their root.

References

Abo Bakr, Hitham, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *The 6th International Conference on Informatics and Systems, INFOS2008*. Cairo University.

Al-Sabbagh, Rania and Roxana Girju. 2010. Mining the Web for the Induction of a Dialectical Arabic Lexicon. In Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *LREC*. European Language Resources Association.

Altantawy, Mohamed, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC. Valletta, Malta.*

Altantawy, Mohamed, Nizar Habash, and Owen Rambow. 2011. Fast Yet Rich Morphological Analysis. In *Proceedings of the 9th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2011)*, Blois, France.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Catalog No.: LDC2002L49.

Chiang, David, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic Dialects. In *Proceedings of the European Chapter of ACL (EACL)*.

Dhouib, Elmoncef. 2007. *El Makki w-Zakiyya*. Publishing House Manshuwrat Manara, Tunis, Tunisia.

Eskander, Ramy, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 573–580, Ann Arbor, Michigan.

Habash, Nizar and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia, July. Association for Computational Linguistics.

Habash, Nizar, Owen Rambow, and George Kiraz. 2005. Morphological Analysis and Generation for Arabic Dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24, Ann Arbor, Michigan.

Habash, Nizar, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In van den Bosch, A. and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., R. Eskander, and A. Hawwari. 2012a. A Morphological Analyzer for Egyptian Arabic. In *NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9.

Habash, Nizar, Mona Diab, and Owen Rabmow. 2012b. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Habash, Nizar, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Habash, Nizar. 2007. Arabic Morphological Representations for Machine Translation. In van den Bosch, A. and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, Nizar. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Hamdi, Ahmed, Rahma Boujelbane, Nizar Habash, and Alexis Nasr. 2013. Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique profonde. In *In proceedings of Traitement Automatique du Langage Naturel (TALN 2013)*.

Kiraz, George Anton. 2000. Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics*, 26(1):77–105, March.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004a. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Maamouri, Mohamed, Tim Buckwalter, and Christopher Cieri. 2004b. Dialectal Arabic Telephone Speech Corpus: Principles, Tool design, and Transcription Conventions. In *NEMLAR International Conference on Arabic Language Resources and Tools*.

Mohamed, Emad, Behrang Mohit, and Kemal Oflazer. 2012. Transforming standard arabic to colloquial arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–180, Jeju Island, Korea, July. Association for Computational Linguistics.

Riesa, Jason and David Yarowsky. 2006. Minimally Supervised Morphological Segmentation with Applications to Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA06)*, pages 185–192, Cambridge,MA.

Salloum, Wael and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Salloum, Wael and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Sawaf, Hassan. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, Denver, Colorado.

Smrž, Otakar. 2007. *Functional Arabic Morphology. Formal System and Implementation*. Ph.D. thesis, Charles University in Prague, Prague, Czech Republic.

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation of Arabic Dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada.